Supplementary Online Material

Mosaic Copy Number Variation in Human Neurons

Michael J. McConnell[1,2,7,8,9], Michael R. Lindberg[7], Kristen J. Brennand[1,10], Julia C. Piper[1,2,11], Thierry Voet[3,4], Chris Cowing-Zitron[1], Svetlana Shumilina[7], Roger S. Lasken[5,6], Joris Vermeesch[3], Ira M. Hall[7,9*], and Fred H. Gage[1*]

1. Laboratory of Genetics, Salk Institute for Biological Studies, La Jolla, CA 92037
2. Crick-Jacobs Center for Theoretical and Computational Biology, Salk Institute for Biological Studies, La Jolla, CA 92037
3. Center for Human Genetics, K.U. Leuven, Leuven, Belgium
4. Wellcome Trust Sanger Institute, Cambridge, UK
5. J. Craig Venter Institute, San Diego, CA 92121
6. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093
7. Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22908
8. Center for Brain Immunology and Glia, University of Virginia, Charlottesville, VA 22908
9. Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908
10. Present Address: Icahn School of Medicine at Mount Sinai, New York, NY 10029
11. Present Address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

* Correspondence to Ira M. Hall (irahall@virginia.edu) and Fred H. Gage (gage@salk.edu)

Contents:

Methods

Human cell culture

The euploid human fibroblasts, human induced pluripotent stem cell (hiPSC)-derived neural progenitor cells (NPCs), and hiPSC-derived neurons used in this study were parallel cultures of the neurotypic control lines reported previously in Brennand et al. (*1*). Reagents were purchased from Life Technologies and their subsidiaries (San Diego, CA) unless noted otherwise. Human fibroblasts from AG09319 (referred to as "D" herein), AG09429 (referred to as "C" herein), AG03651 (referred to as "E" herein), and GM01920 (trisomic male) were obtained from the Coriell Institute (Camden, NJ) and grown in DMEM with Glutamax supplemented with 15% FBS (Atlanta Biologicals, Atlanta, GA).

Briefly, reprogramming was initiated using a cocktail of 5 tetracycline-inducible lentivirus (LV) vectors expressing human *OCT4, SOX2, c-MYC, KLF4,* and *LIN28* cDNAs. Human fibroblasts were infected every day for five days. Following infection, fibroblasts were plated on a mouse embryonic fibroblast (MEF) feeder layer and switched to HUES media (KO-DMEM, 10% KO-Serum Replacement, 10% Plasminate, 1x Glutamax, 1x NEAA, 1x 2-mercaptoethanol and 20 ng/ml bFGF2 (Peprotech, Rocky Hill, NJ), supplemented with 1ug/mL doxycycline (Sigma, St. Louis, MO). Successful reprogramming was confirmed by human embryonic stem (ES) cell-like morphology, by expansion and maintenance of a euploid karyotype beyond 15 passages, by expression of endogenous pluripotency genes (e.g. *OCT4, SOX2, NANOG, REX1,* and *CRIPTO* mRNA) and proteins (OCT, SOX2, NANOG, and TRA-1-60), and, importantly, by repression of LV genes in the absence of doxycycline.

Karyotypically normal hiPSCs were used to derive NPCs. hiPSCs were enzymatically dissociated from the MEF feeder layer using Collagenase type IV and grown in suspension as embryoid bodies (EBs) in N2/B27 media (DMEM/F12-Glutamax, 1X N2, 1XB27). After 1 week, EBs were transferred onto polyornithine (PORN)/laminin-coated plates in N2 media containing 1 µg/ml laminin. After an additional week of differentiation, neural rosettes formed; these were manually dissected, dissociated, and plated onto PORN/laminin-coated plates in NPC media (N2/B27 media with 1 µg/ml laminin and 20 ng/ml FGF-2) to expand NPCs. hiPSC-derived NPCs (passages 7 and 8) were differentiated into neurons in neural differentiation media (DMEM/F12-Glutamax, 1X B27-RA, 1X N2 with 20 ng/ml BDNF, 20 ng/ml GDNF (Peprotech), 1 mm dibutyrl-cyclicAMP (Sigma), 200 nm ascorbic
acid (Sigma)) for 7 weeks.

Karyotyping and FISH were performed by WiCell Cytogenetics (Madison, WI). FISH probes for ChrX (Kallman probe set) were obtained from Abbott Laboratories (Abbott Park, IL). The ChrX p arm probe is specific for ChrXp22.3. The centromeric Chr20 probe is from Cytocell (Cambridge, UK). The Chr20 q arm probe is specific for Chr20q21 (RPCI-11 702M8-552, Empire Genomics, Buffalo, NY).

Isolation of single cells

Confluent fibroblast cultures (passage 7 – 10) were serum-starved for 72 hours; G1 arrest was confirmed on a subset of this population using flow cytometry. NPCs (passages 9 and 10) were refractory to serum starvation; therefore, possible analysis of some S or G2 cells cannot be excluded. Single cells were picked by hand using a micropipette ("the Stripper") and 75 uM glass pipettes (Origio Midatlantic Devices, Mt. Laurel, NJ).

Five-week-old hiPSC-derived neuronal cultures were infected twice with a LV construct (*1*) where GFP expression is driven by a synapsin promoter (Syn::GFP). Two weeks later, cells were dissociated using TrypLE and counterstained with 10 ug/mL propidium iodide (PI). GFP-positive, PI-negative cells were isolated via fluorescence activated cell sorting (FACS) on a FACS Aria II (BD Biosciences, San Jose, CA). Neurons were sorted into DMEM with 10% FBS and 10% DMSO and then frozen at -80C in Styrofoam. Frozen vials of hiPSC-derived neurons were thawed and individual cells isolated manually as before (*2*).

Single cells were lysed and genomic DNA amplified via multiple displacement amplification (MDA) using phi29 polymerase (Genomiphi V2, GE Healthcare, Piscataway, NJ) as described (*2*). MDA products (5 ng) were examined for even amplification (e.g., +/- 5% of the Ct for 5 ng bulk genomic DNA) using qPCR (Applied Biosystems, San Diego, CA). To test for even amplification, we used a 10 locus subset of the 47 single copy loci used in Hosono et al. (*3*) (here, Chr1p, Chr2p, Chr3q, Chr7p, Chr10p, Chr11p, Chr14q, Chr17q, Chr19p, and Chr21q), similar to the approach employed previously for MDA QC (*4, 5*).

Detection of copy number variation (CNV) from microarray data

MDA products passing qPCR quality control (QC) measures were analyzed on Affymetrix 250K NSP chips (Affymetrix, San Jose, CA). Partek Genomics Suite Software (version 6.6 beta, Partek, St. Louis, MO) was used to calculate predicted copy numbers for each probe set intensity. A custom copy number model composed of 161 MDA single cell experiments (from this and other studies) was generated to perform quantile normalization of the calculated copy numbers. The background-adjusted values were then subjected to GC correction in windows of 10 Mb, and artifact-prone probes were removed according to Pugh et al. (*6*). We then performed smoothing by taking the median copy number value in non-overlapping genomic windows composed of 100 probes. On average, each 100-probe bin corresponds to 666 Kb of genome sequence. At this stage we also excluded 6 of 107 samples that had excessively "noisy" copy number profiles, defined as having a median absolute deviation (MAD) greater than 0.7. To detect CNVs, we used the circularly binary segmentation (CBS) algorithm (*7*) from the DNAcopy package in R, with the following parameters: *alpha=0.001, undo.splits="sdundo", undo.SD=1*. We defined CNVs as segments composed of 10 or more contiguous genomic windows whose copy number value differed from the dataset's median copy number by at least 1 MAD. We did not attempt to detect CNVs on the Y chromosome.

Isolation of post-mortem neuronal nuclei

Postmortem human frontal cortex tissues from UMB#5125 (a neurotypic 24-year-old female, 9 hour post-mortem interval), UMB#1846 (a neurotypic 20-year-old female, 9 hour post-mortem interval) and UMB#1583 (a neurotypic 26-year-old male, 18 hour post-mortem interval) were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland. Tissue samples were placed in nuclear isolation medium [(NIM) 25 mM KCl, 5 mM MgCl2, 10 mM Tris-Cl, 250 mM sucrose, 1mM dithiothreitol (DTT), and 1X protease inhibitor cocktail (Roche)] and homogenized with a polytron tissue homogenizer (Kinematica, Inc., Bohemia, NY). Homogenized tissue was supplemented with 0.1% TritonX-100, and further processed using a dounce homogenizer. Samples were centrifuged (1,000xg, 8 min) and the pellet was resuspended in 10:5:1 NIM:Iodixanol (Sigma):OptiPrep Diluent for Nuclei [(ODN) 150 mM KCl, 30 mM MgCl2, 60 mM Tris-Cl, 250 mM sucrose)]. Samples were layered onto a 29% Iodixanol in ODN cushion using a 1 mL syringe and centrifuged (10,300xg, 20 min, 4°C) in a Beckman L8-M ultracentrifuge with SW55 Ti rotor. Pellets were resuspended in nuclei storage buffer [(NSB), 5 mM MgCl2, 50 mM TrisCl, 166 mM sucrose, 1 mM DTT, and 1X protease inhibitor cocktail]. Free nuclei and purity were confirmed visually by microscope.

Neuronal nuclei were purified from bulk brain nuclei using NeuN immunostaining (*8, 9*). Immunostaining was performed for 1 hour at 4°C with gentle agitation in PBS containing 5 ug/mL (1:2000) AF488-conjugated NeuN (Chemicon, Billerica, MA). Nuclei were then stained for DNA content with 10 ug/mL DAPI and analyzed by FACS. Single cells from the NeuN and

DAPI positive population were sorted into 96 well plates alongside 1 water control per row. For benchmarking experiments, trisomic male fibroblasts were similarly sorted into 96 well plates based on size and propidium iodide exclusion.

Single cell sequencing

Isolated single neuronal nuclei (or whole fibroblast cells) were lysed and amplified using the WGA4 GenomePlex Single Cell Whole Genome Amplification Kit (Sigma), using 15 cycles of PCR amplification. In the case of replicate fibroblast experiments, reactions were split into two after 8 cycles of PCR and then subjected to an additional 8 cycles. Subsequent WGA4 products were purified with Qiagen mini-elute columns (Qiagen, Germantown, MD). Illumina-compatible sequencing libraries were constructed using the Nextera Sample Prep (Epicentre Biotechnologies, Madison, WI and Illumina, San Diego, CA) according to the manufacturer's protocol, with the modification that we used a 1:200 dilution of the "transposome" enzyme complex in the "tagmentation" reaction (which helps control the fragment size distribution in single cell reactions). Tagmented DNA fragments were purified with mini-elute columns (Qiagen) and subjected to 12-15 cycles of PCR, during which barcodes were added to each library to facilitate pooled sequencing. The resulting barcoded sequencing libraries were purified with mini-elute columns (Qiagen). Each library was run on a 2% Low Range Ultra Agarose gel (Bio-Rad, Hercules, CA) with TAE and stained with SYBR Gold (Invitrogen) for 10-40 minutes. The 200-600 bp size fraction was isolated by gel extraction and purified with the QIAquick kit (Qiagen). Frontal cortex neuron libraries were sequenced with paired-end reads on an Illumina GAIIx with 38-39 bp reads and fibroblasts libraries were sequenced with single-end reads on an Illumina MiSeq with a read length of 59 bp.

Detection of CNVs from single cell sequencing data

Copy number was assessed in dynamically sized genomic windows containing 500 Kb of uniquely mappable DNA sequence, as defined by the *wgEncodeCrgMapabilityAlign40mer* track from the UCSC Genome Browser (*10*). The mean absolute window size was 687 Kb. Paired-end reads were aligned to the human genome (NCBI Build 37) using BWA (version 0.5.10) with default settings (*11*), and duplicates were removed using *MarkDuplicates* from the Picard software suite (http://picard.sourceforge.net/). Read-depth analysis was performed essentially as described previously (*12-14*). Read-depth was assessed using *coverageBed* from the BEDTools software suite (*15*). Since Illumina sequence coverage is known to vary due to GC content, to obtain the predicted copy number of each genomic window we divided the read-depth of that window by the genome-wide median read-depth of all windows with similar GC content, as measured in 1-3% intervals, then multiplied by 2. CNVs were identified using the CBS algorithm (*7*) with the aforementioned parameters. We defined CNVs as segments composed of 5 or more contiguous genomic windows whose copy number value differed from the dataset's median copy number by at least 2 MADs. CNVs were not called on the Y chromosome. For putative CNVs on the X chromosome in the male sample, the median and MAD of the X chromosome were used to filter CNV calls.

The final CNV callset only includes datasets that passed the following QC criteria: 1) the dataset contained more than 5x10^5 reads following duplicate removal; 2) the median absolute deviation of predicted copy number values in autosomal genomic windows was not more than 0.35; and 3) the dataset had a confidence score of at least 0.85, as defined below. In total, 110 of 208 datasets passed all of these QC filters.

The confidence score, *S,* is a measure of the extent to which a given dataset adheres to the expectation of integer-like copy number measurements. The rationale for this QC measure is that we are using a digital technology (DNA sequencing) to measure copy number in single cells, and thus there is a strong expectation that copy number profiles should display approximately integer values. Non-integer copy number values may potentially occur due to regional variability in DNA amplification efficiency or flow-sorting errors that result in multiple nuclei being deposited into a single well.

Confidence Score, *S*: $$S = 1 - 2\,\frac{\sum_{i=0}^{n} \min(\lceil C_i \rceil - C_i, C_i - \lfloor C_i \rfloor)}{n}$$

C: the median predicted copy number of a given genomic interval (*i*) after copy number segmentation
n: the total number of genomic windows in the dataset

This score is the average distance between the predicted absolute copy number of each genomic segment in the dataset (as defined by the CBS algorithm) to the nearest integer value. This computed average is then multiplied by a factor of two in order to compare the actual distances to the worst-case distance (0.5) for every interval. This actual to worst-case ratio is then subtracted from 1 to yield a score between 0 and 1, where the more digital the dataset, the closer this score is to 1. Therefore, a dataset with a score of 0.85 or higher is very close to the assumed model of a dataset being primarily composed of integer copy number values.

Cluster analysis comparison of replicate sequencing and microarray data

Both sequencing and microarray analyses were performed for 7 of the hiPSC-derived neurons. To enable straightforward comparison of these two data types across the same genomic intervals, for this analysis we aggregated SNP array data in the same genomic windows as sequencing data (rather than 100-probe windows). The mean window size is 687 Kb, and the windows contained a mean of 57.8 probes (median 58). Only 6 of the windows had zero probes and these were assumed to have a copy number of 2. The microarray data processed in this manner are somewhat more noisy than those analyzed with a 100-probe window, but overall data quality is similar. To assess the concordance between sequencing and microarray methods, the raw per-window copy number values of these 14 datasets were subjected to unsupervised clustering using the pvclust package (http://www.is.titech.ac.jp/~shimo/prog/) in R, using default parameters: *distance=correlation, linkage=average*.

Enrichment analyses

For enrichment analyses we used the BITS algorithm (*16*) to count the observed number of overlaps between CNVs and various genome annotations. The fragile sites track was obtained from Fungtammasan et al. (*17*); all other tracks were downloaded from the UCSC Genome Browser (*10*). For these analyses we used CNVs less than 20 Mb in size, which reduces the total callset from 148 to 133. We then conducted Monte-Carlo simulations to find the expected number of intersections by shuffling both the CNVs and annotation track 1,000 times. The log2 enrichment ratio was caclulated as the observed number of overlaps divided by the median (or mean if the median was 0) number of intersections observed in simulations. Analyses of telomeric enrichment were performed in a similar way; however, only the CNVs were shuffled for the 1,000 iterations.

Estimating the false discovery rate (FDR) of CNV detection by read-depth analysis

To estimate the FDR for CNV detection by read-depth analysis, we performed Monte-Carlo simulations in which the relative order of genomic windows was shuffled 1,000 times for each dataset. Shuffled datasets were subjected to copy number segmentation and filtering exactly as for real data, with the caveat that we excluded the X and Y chromosomes from these analyses to avoid sex-related effects. The FDR was calculated as the mean number of CNVs detected in simulated data, adjusted for the exclusion of sex chromosomes (based on their size). This FDR estimation strategy specifically measures the specificity of CNV detection with respect to random sources of noise; however, it does not account for potential systematic or regional effects and therefore should be considered a lower bound.

Estimating the false negative rate (FNR) of CNV detection by read-depth analysis

It is difficult to estimate the FNR because our CNV size detection limits (~3.4 Mb) greatly exceed the size of known germline CNVs; therefore we do not have access to a set of true CNVs with which to measure sensitivity. However, for the 41 cells derived from male individual 1583, we expected to detect the X and Y chromosomes as single copy "aberrations" relative to autosomes. We exploited this feature to develop a simulation-based approach to measure FNR in these 41 datasets. For example, to simulate a single deletion comprising 5 genomic windows, we randomly selected 5 contiguous genomic windows from the X chromosome, extracted their predicted copy number values, and used these values to replace the copy number values of 5 contiguous windows from a randomly selected autosomal location. To simulate duplications, we used a similar approach but, instead of replacing the 5 autosomal copy number values, we simply added the autosomal values to the values extracted from the X chromosome. The resulting simulated dataset was then subjected to copy number segmentation and CNV filtering precisely as for the real data. To calculate the FNR for CNVs of a given size (e.g., 5 windows) in a given dataset, we simulated 1,000 CNVs of that size and assessed the fraction of simulations in which we detected the synthetic CNV. Detection was defined as a reciprocal overlap of 50% between the simulated and detected genomic segment. The average of the deletion and duplication detection rates were reported as a composite FNR rate.

References

1. K. J. Brennand *et al.*, Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473**, 221 (May 12, 2011).
2. E. Vanneste *et al.*, Chromosome instability is common in human cleavage-stage embryos. *Nat Med* **15**, 577 (May, 2009).
3. S. Hosono *et al.*, Unbiased whole-genome amplification directly from clinical samples. *Genome Research* **13**, 954 (May, 2003).
4. Y. Li *et al.*, Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience* **1**, 12 (2012).

5. X. Xu *et al.*, Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886 (Mar 2, 2012).

6. T. J. Pugh *et al.*, Impact of whole genome amplification on analysis of copy number variants. *Nucleic acids research* **36**, e80 (Aug, 2008).

7. A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557 (Oct, 2004).

8. J. W. Westra *et al.*, Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. *The Journal of comparative neurology* **518**, 3981 (Oct 1, 2010).

9. K. L. Spalding, R. D. Bhardwaj, B. A. Buchholz, H. Druid, J. Frisen, Retrospective birth dating of cells in humans. *Cell* **122**, 133 (Jul 15, 2005).

10. L. R. Meyer *et al.*, The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research* **41**, D64 (Jan, 2013).

11. H. Li, R. Durbin, Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, (May 18, 2009).

12. A. R. Quinlan *et al.*, Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research* **20**, 623 (May, 2010).

13. A. R. Quinlan *et al.*, Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. *Cell stem cell* **9**, 366 (Oct 4, 2011).

14. A. Malhotra *et al.*, Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Research* **23**, 762 (May, 2013).

15. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841 (Mar 15, 2010).

16. R. M. Layer, K. Skadron, G. Robins, I. M. Hall, A. R. Quinlan, Binary Interval Search: a scalable algorithm for counting interval intersections. *Bioinformatics* **29**, 1 (Jan 1, 2013).

17. A. Fungtammasan, E. Walsh, F. Chiaromonte, K. A. Eckert, K. D. Makova, A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome Research* **22**, 993 (Jun, 2012).

Figure S1

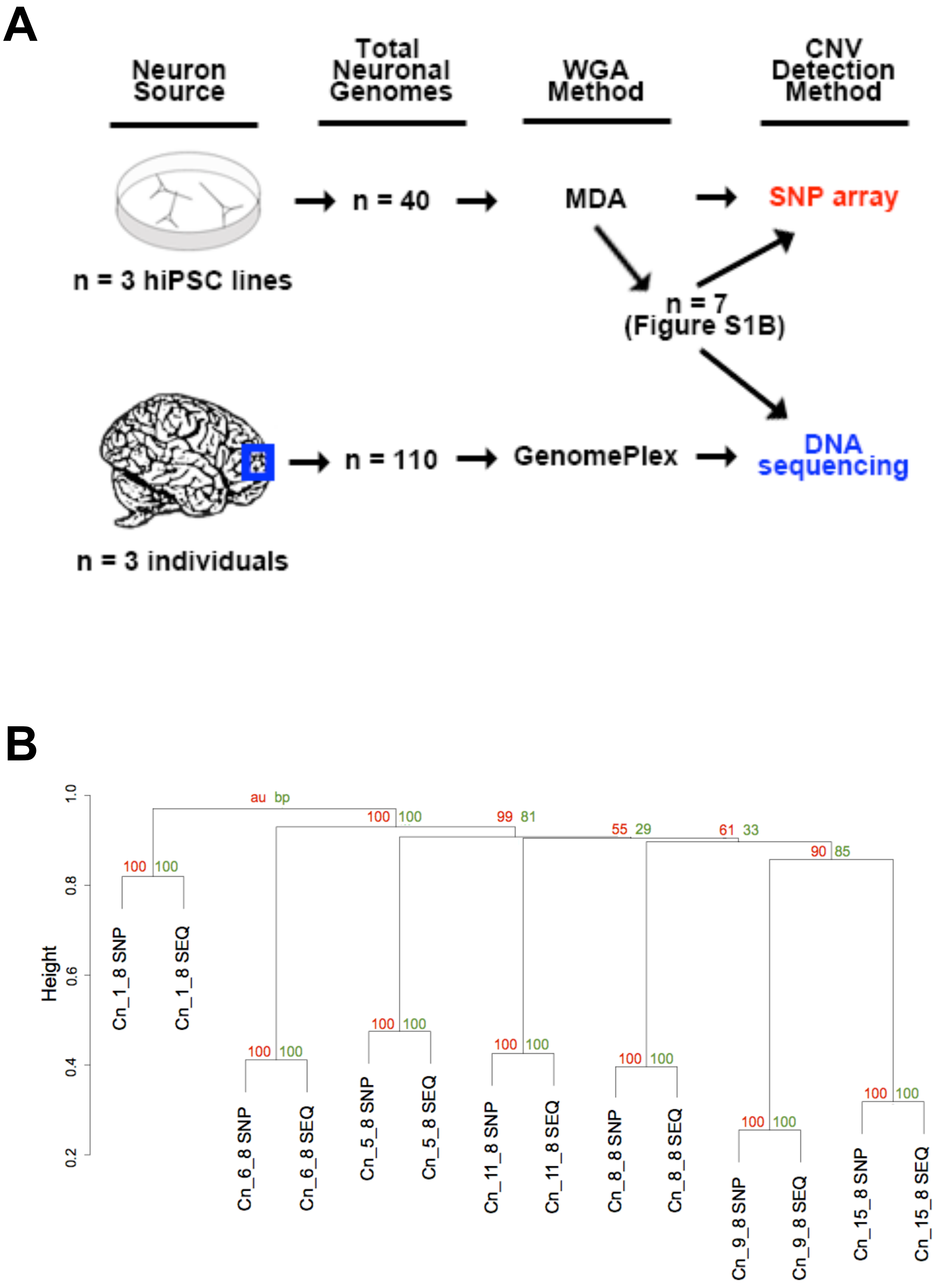


Figure S1. Single cell analysis by SNP array and DNA sequencing. (A) Summary of the single cell approaches used in the study. **(B)** A cluster dendrogram shows concordance in copy number profiles for seven neurons mapped by SNP array hybridization intensity ("SNP") or sequencing read depth ("SEQ"). Numbers at tree nodes reflect the significance values reported by the R pvclust package for bootstrap resampling (1000 iterations) and can be interpreted as the percentage of simulated trees with the observed topology.

Figure S2

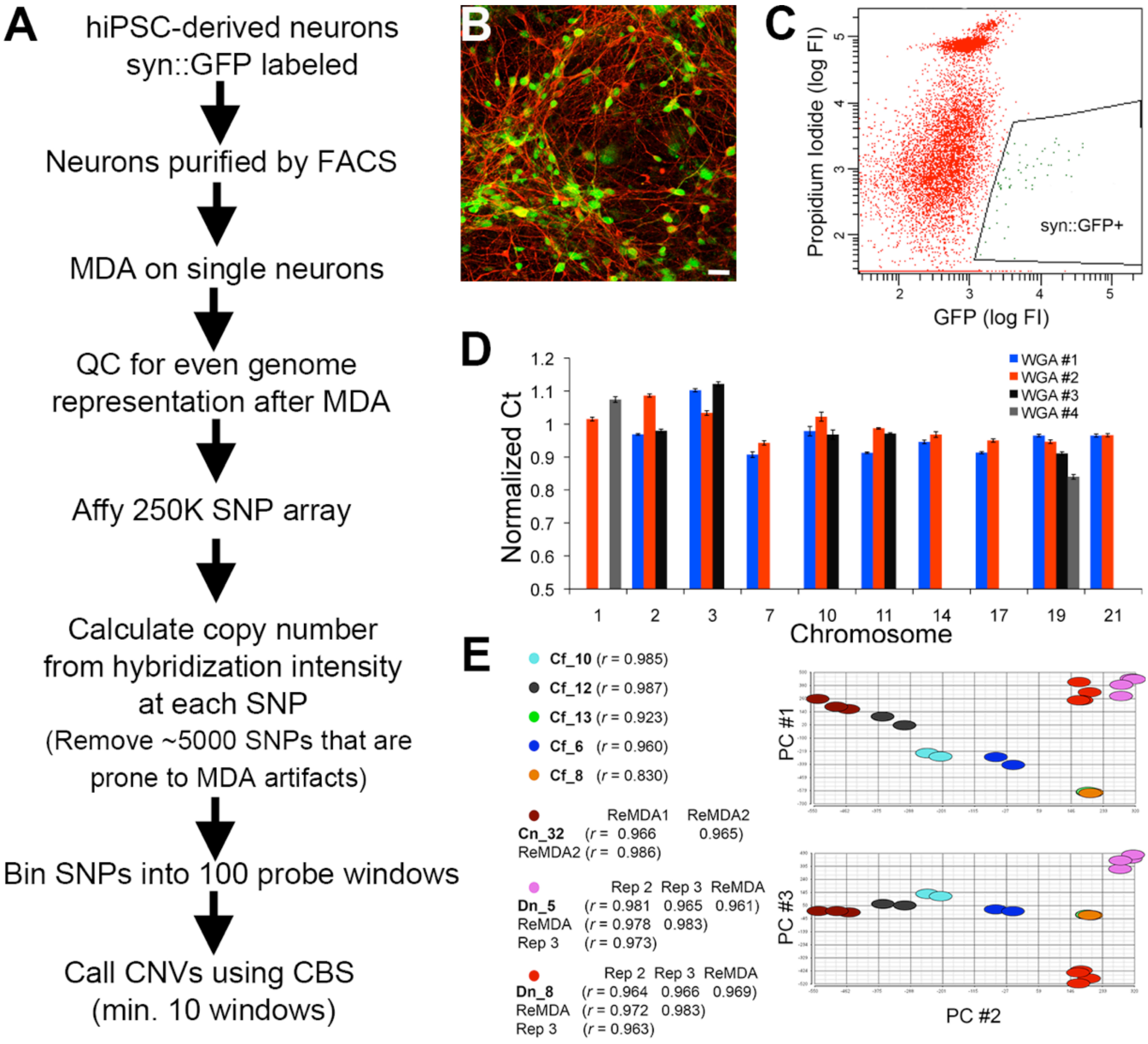


Figure S2. Single cell analysis of hiPSC-derived neurons. (A) Flow chart of the protocol. **(B)** The synapsin::GFP (syn::GFP) reporter faithfully identifies neurons. Extensive co-localization of GFP (green) and MAP2ab immunostaining (red) is observed in syn::GFP-infected hiPSC-derived neurons. Scale bar = 20 um. **(C)** FACS is used to purify hiPSC-derived neurons. After dissociation, most cells lose membrane integrity and are permeable to propidium iodide (PI). The GFP-positive neurons were sorted from those cells that continued to exclude PI. **(D)** Quantitative PCR of 10 loci on different chromosomes was used to identify high quality single cell genome amplifications. Two high-quality MDA reactions (blue, red) are shown; 9 or 10 loci are represented comparably to that of bulk genomic DNA. Two poor-quality MDA reactions (gray and black) are also shown. **(E)** Principal component analysis of technical replicates from 3 hiPSC-derived neurons and 5 fibroblasts. Pearson correlation coefficients (*r*) for replicates are indicated in the key.

Figure S3

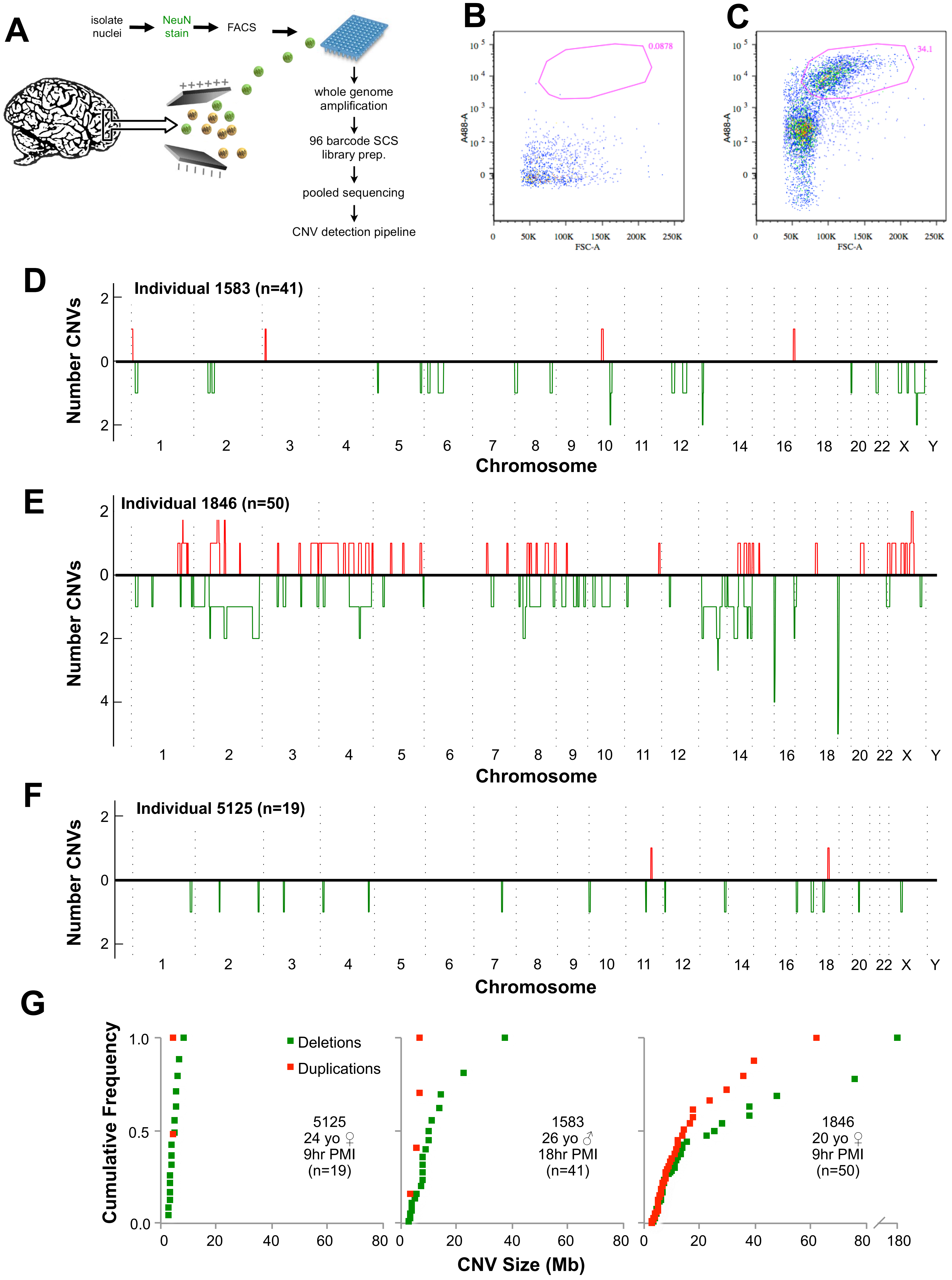
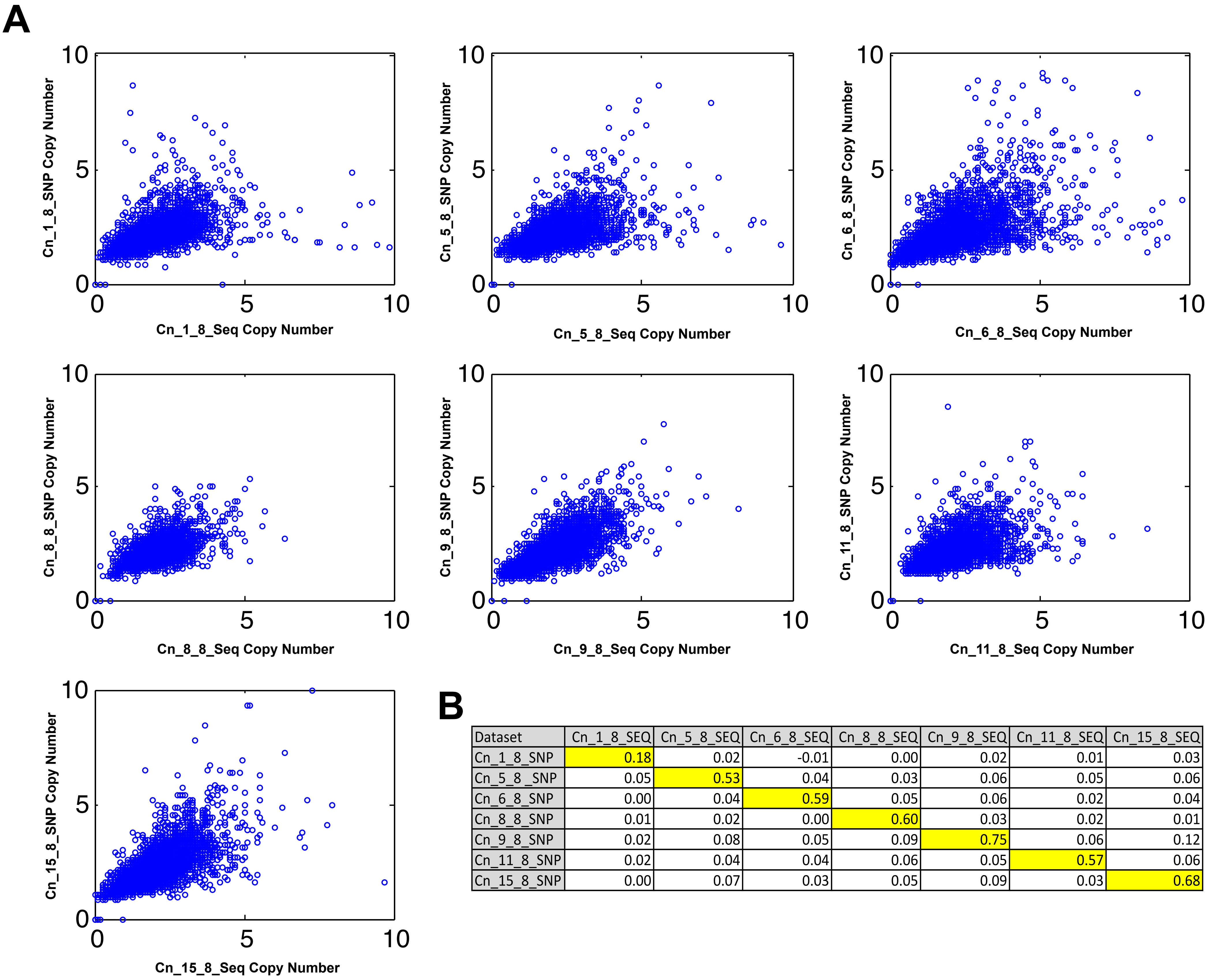


Figure S3. Single cell analysis of FCTX neurons. (A) Flow chart of the protocol. **(B, C)** FACS-based identification of large nuclei that stain positive for NeuN (C), relative to unstained controls (B). Sorted nuclei are gated from the pink circle. **(D, E, F)** Summary of duplications and deletions for each individual (number indicated) plotted as in Fig. 3B. The y-axis values represents the number of times each genomic interval was deleted (below in green) or duplicated (above in red). **(G)** Cumulative frequency of CNV sizes found per individual (deletions in green, duplications in red).

Figure S4



B

Dataset	Cn_1_8_SEQ	Cn_5_8_SEQ	Cn_6_8_SEQ	Cn_8_8_SEQ	Cn_9_8_SEQ	Cn_11_8_SEQ	Cn_15_8_SEQ
Cn_1_8_SNP	0.18	0.02	-0.01	0.00	0.02	0.01	0.03
Cn_5_8_SNP	0.05	0.53	0.04	0.03	0.06	0.05	0.06
Cn_6_8_SNP	0.00	0.04	0.59	0.05	0.06	0.02	0.04
Cn_8_8_SNP	0.01	0.02	0.00	0.60	0.03	0.02	0.01
Cn_9_8_SNP	0.02	0.08	0.05	0.09	0.75	0.06	0.12
Cn_11_8_SNP	0.02	0.04	0.04	0.06	0.05	0.57	0.06
Cn_15_8_SNP	0.00	0.07	0.03	0.05	0.09	0.03	0.68

Figure S4. Concordance between SNP-array and DNA sequencing. (A) Scatter plots comparing raw copy number values between the seven neurons subjected to MDA-based whole-genome amplification followed by both SNP-array analysis ("SNP") and DNA sequencing ("SEQ"). Copy number values were directly compared using the same ~687Kb windows used to measure read-depth (see methods). **(B)** Correlation matrix reporting pairwise Pearson correlation coefficients for every "SNP" and "SEQ" combination. Note that replicate SNP/SEQ experiments have dramatically larger correlation coefficients than non-replicate combinations.

Figure S5: see separate file

Figure S5. Genome-wide copy number profiles for all SNP-array datasets. Plots showing genome-wide copy number profiles for all single cells analyzed by MDA plus SNP-array. In each plot, the raw predicted copy number values for each individual genomic bin are shown in blue, and the copy number profiles obtained by circular binary segmentation are shown in orange. The gray dotted lines show 1 and 2 median absolute deviations (MADs) from the median copy number of each dataset, which represent thresholds used for CNV filtering. Note, final CNV calls were required to comprise 10 consecutive bins and to differ from the dataset median by 1 MAD.

Figure S6

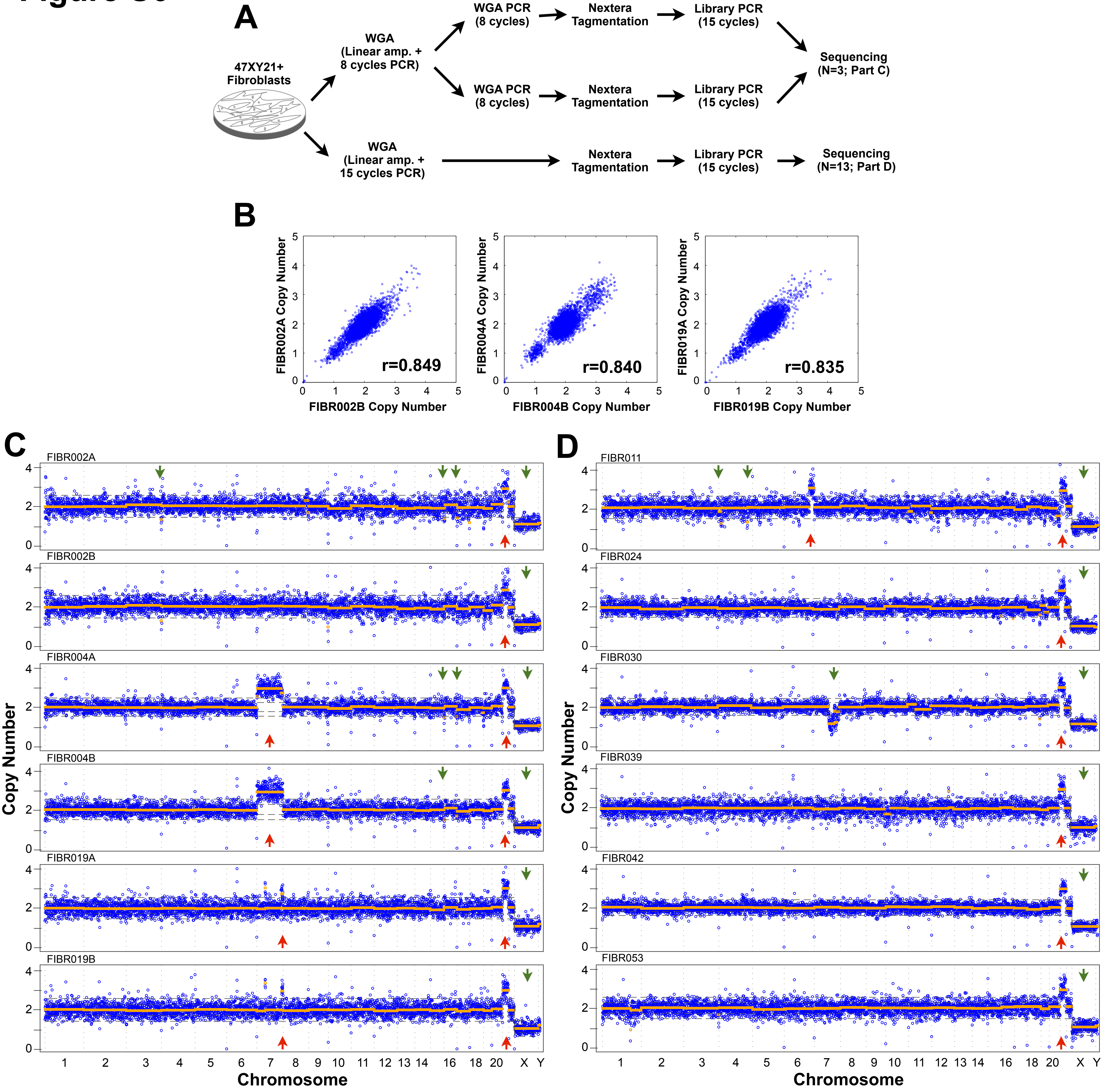


Figure S6 Identification of CNVs in male trisomy 21 fibroblasts using single cell sequencing. (A) Flow chart of the single fibroblast sequencing experiment. The top section shows the protocol for sequencing 3 cells in replicate (corresponding to B & C), accomplished by splitting each sample after 8 cycles of the whole-genome amplification (WGA) PCR step. The standard protocol used for neurons and the 13 single fibroblasts (corresponding to D) is shown in the bottom section. QC filtering was performed exactly as for neurons. **(B)** Scatter plots comparing concordance between replicate experiments, where each data point represents the predicted copy number of a single genomic window. (**C**) Genome-wide copy number profiles of the three replicate fibroblasts. Read-depth analysis, copy number segmentation and CNV filtering were performed exactly as for neurons. These plots follow the conventions of Figure 3A. Blue dots represent the predicted copy number (Y-axis) of each individual genomic window, and orange lines show the results of copy number segmentation. Dotted gray lines show 1 and 2 MADs from the median copy number of each dataset. Reported CNVs comprise five or more consecutive bins and exceed two MADs. Arrows indicate CNV calls that passed filtering criteria (deletions in green and duplications in red). (**D**) Genome-wide copy number profiles of six single fibroblast cells, shown following the conventions of Figure 3A and outlined for part C above.

Figure S7: see separate file

Figure S7. Genome-wide copy number profiles for all single cell sequencing datasets. Plots showing the genome-wide copy number profile of all single cells sequencing experiment included in this study. In each plot, the raw predicted copy number values for each individual genomic bin are shown in blue, and the copy number profiles obtained by circular binary segmentation are shown in orange. The gray dotted lines show 1 and 2 MADs from the median copy number of each dataset, which represent thresholds used for CNV filtering. Note that for sequencing experiments, the final CNV calls were required to comprise 5 genomic bins and to exceed 2 MADs.

Figure S8

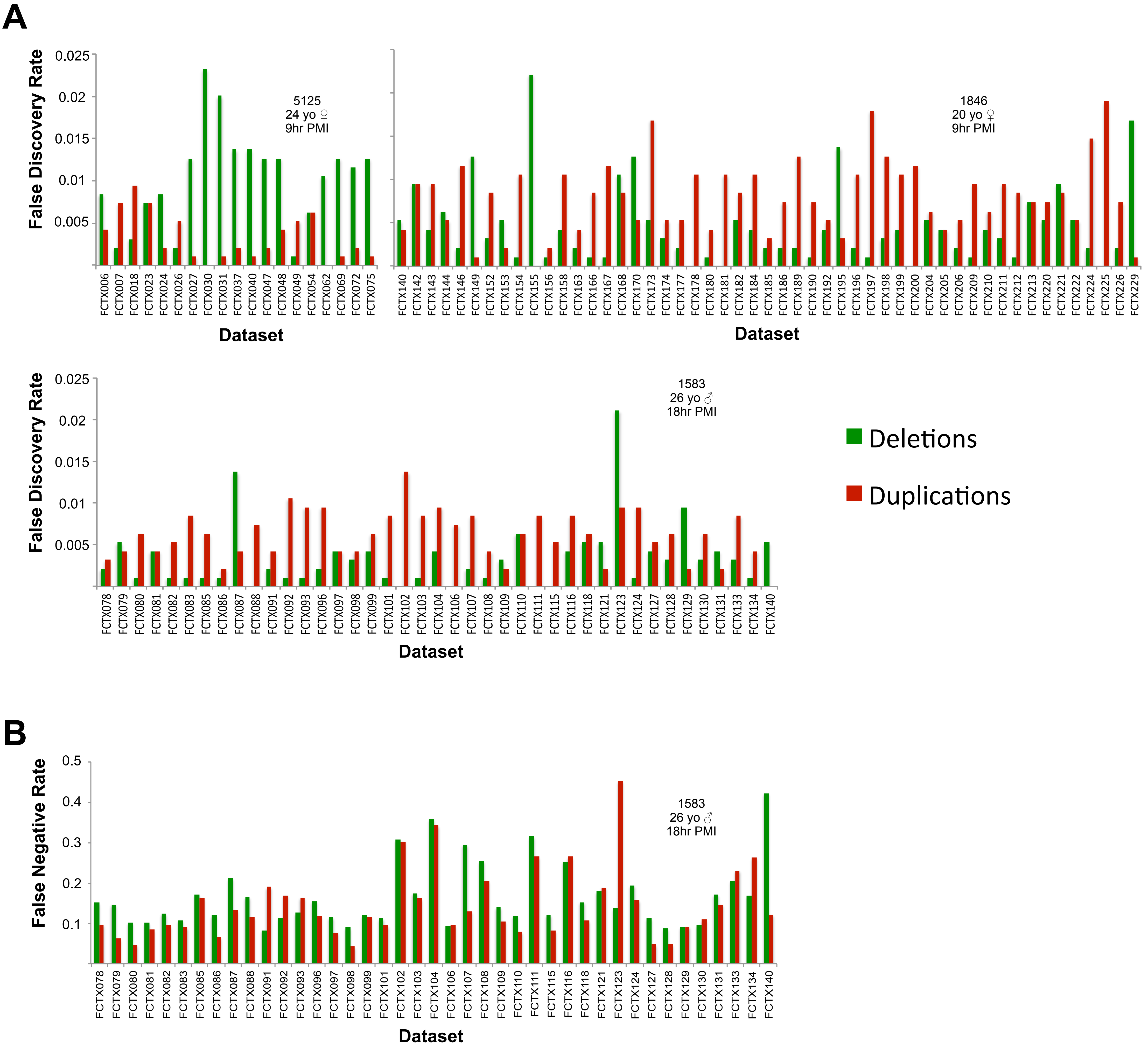


Figure S8. Estimated false discovery rate (FDR) and false negative rate (FNR) for single cell sequencing experiments. In each case, CNVs were identified using precisely the same methods and criteria as for real data (see Methods), and the FDR or FNR shown is the mean value obtained from 1000 simulation experiments. Deletions are shown in green and duplications in red. **(A)** FDR for each dataset, as determined by randomly shuffling copy number values across all autosomal bins and then calling CNVs. **(B)** FNR for all cells derived from the male individual (1583). FNR was calculated by randomly selecting 5 contiguous bins from the X chromosome and either replacing (deletion) or adding (duplication) the copy number values from the these bins at a randomly chosen genomic location.

Figure S9

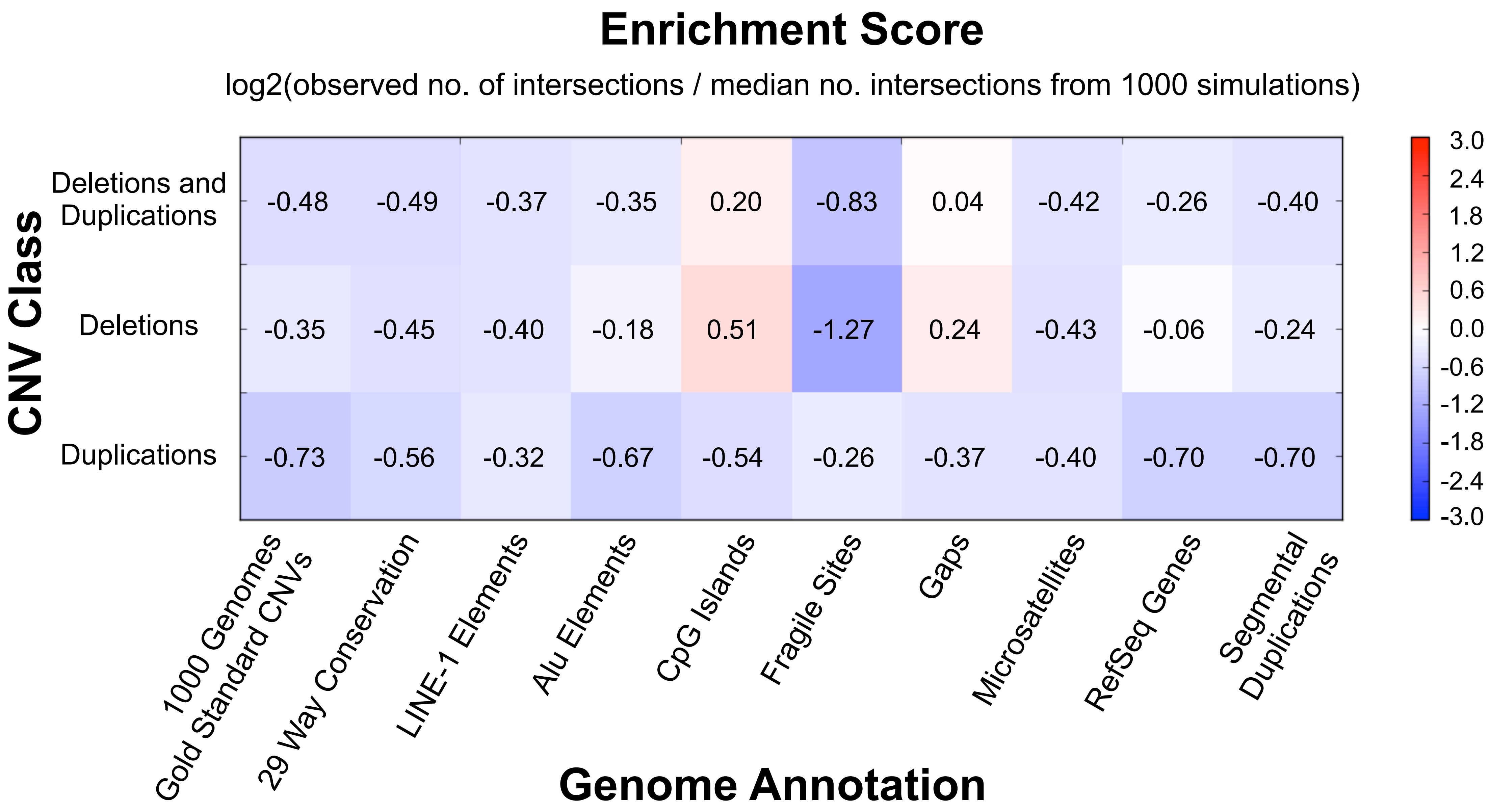


Figure S9. Enrichment of CNV calls at various genome annotations. Monte-Carlo simulations were used to determine whether CNVs identified in post-mortem neurons preferentially overlapped various genomic features. Enrichments are displayed as the log2 ratio of the observed number of intersections between each CNV class (x-axis) and each genome annotation (y-axis), relative to the expected number of random intersections calculated by the simulations. A positive correlation between CNVs and a given annotation will result in a red-colored positive value; an anti-correlation will result in a blue-colored negative value. The highest level of enrichment observed was between deletions and CpG islands, whereas the lowest level of enrichment observed was between deletions and fragile sites.

Figure S10

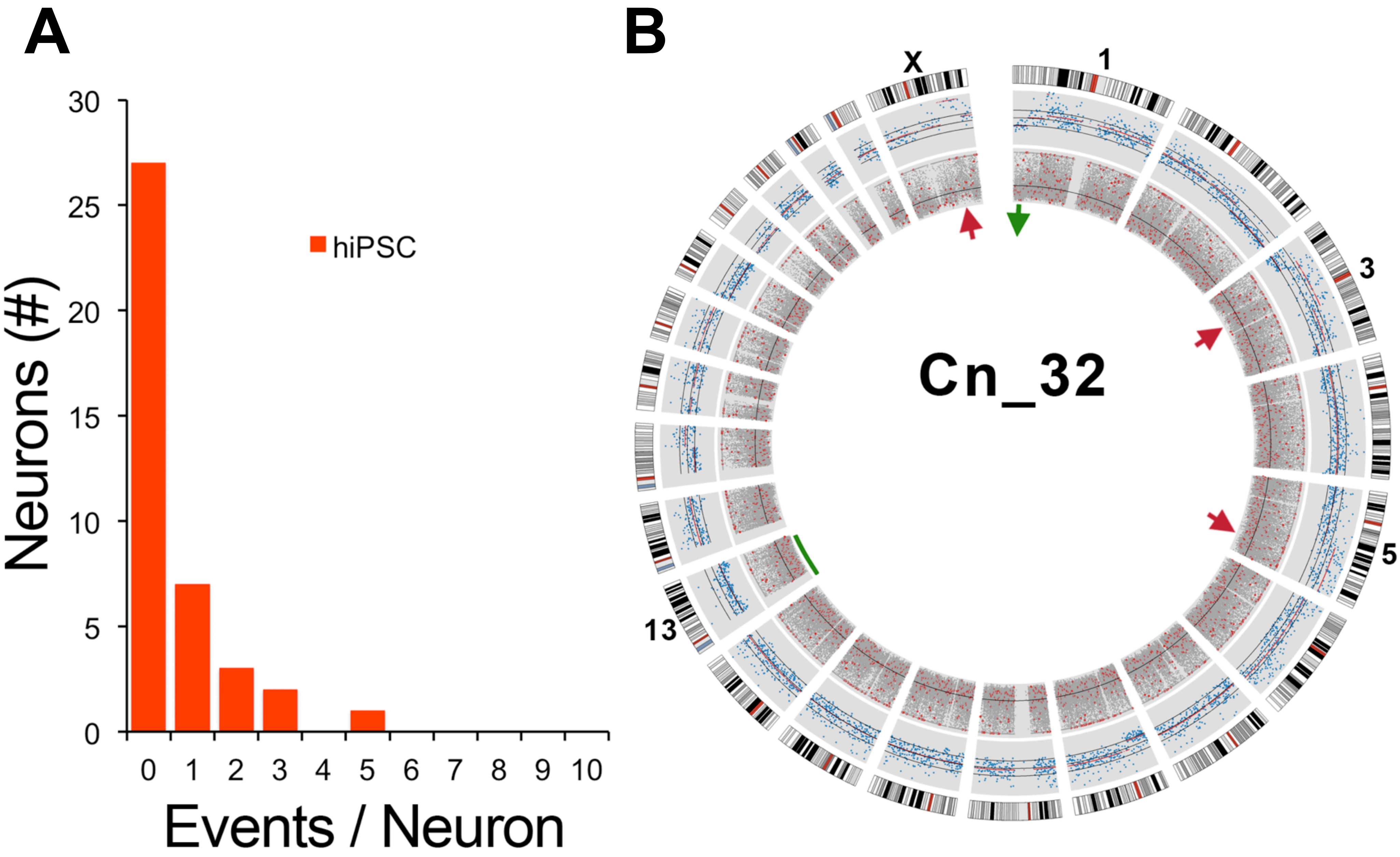


Figure S10. A subset of hiPSC-derived neurons carries multiple CNVs. (A) Most hiPSC-derived neurons have 0 or 1 events. **(B)** One hiPSC-derived neuron had 5 events. A CIRCOS plot shows Cn_32 with duplications on Chr3, Chr6, and ChrX (red arrows), a deletion on Chr1 (green arrow), and aneuploidy for Chr13 (-13, green bar). The innermost ring shows SNP copy number data as in Fig. 1C and D. The adjacent ring (blue markers) shows binned copy number data from which events were called.

Figure S11

A

Frontal Cortex Neurons (110 cells)						
Stringency	**Total CNV Calls**	**Deletions**	**Duplications**	**Cells w/ >=1 CNV**	**Predicted FNR**	**Monosomy X detected**
2 MADs, 5 bins	148	98	50	45 (41%)	17%	41 (100%)
3 MADs, 5 bins	83	50	33	22 (20%)	22%	37 (90%)
2 MADs, 10 bins	73	47	26	26 (24%)	7%	41 (100%)
3 MADs, 10 bins	45	29	16	14 (13%)	15%	37 (90%)

B

Fibroblasts - Single Datasets (13 cells)						
Stringency	**Total CNV Calls**	**Deletions**	**Duplications**	**Cells with >=1 CNV**	**Monosomy X detected**	**Trisomy 21 detected**
2 MADs, 5 bins	7	6	1	4 (31%)	13 (100%)	13 (100%)
3 MADs, 5 bins	3	2	1	2 (16%)	10 (77%)	10 (77%)
2 MADs, 10 bins	5	4	1	4 (31%)	13 (100%)	13 (100%)
3 MADs, 10 bins	3	2	1	2 (16%)	10 (77%)	10 (77%)

C

Fibroblasts - Replicate Datasets (3 cells, 6 datasets)								
Stringency	**Total CNV Calls**	**Deletions**	**Duplications**	**Cells with >=1 CNV**	**Concordant CNVs**	**Discordant CNVs**	**Monosomy X detected**	**Trisomy 21 detected**
2 MADs, 5 bins	7 (11)	2 (5)	5 (6)	3 (100%)	7	4	3 (6)	3 (6)
3 MADs, 5 bins	2 (4)	2 (4)	0 (0)	2 (67%)	3	1	3 (6)	3 (5)
2 MADs, 10 bins	2 (5)	2 (5)	0 (0)	2 (67%)	5	0	3 (6)	3 (6)
3 MADs, 10 bins	2 (4)	2 (4)	0 (0)	2 (67%)	3	1	3 (6)	3 (5)

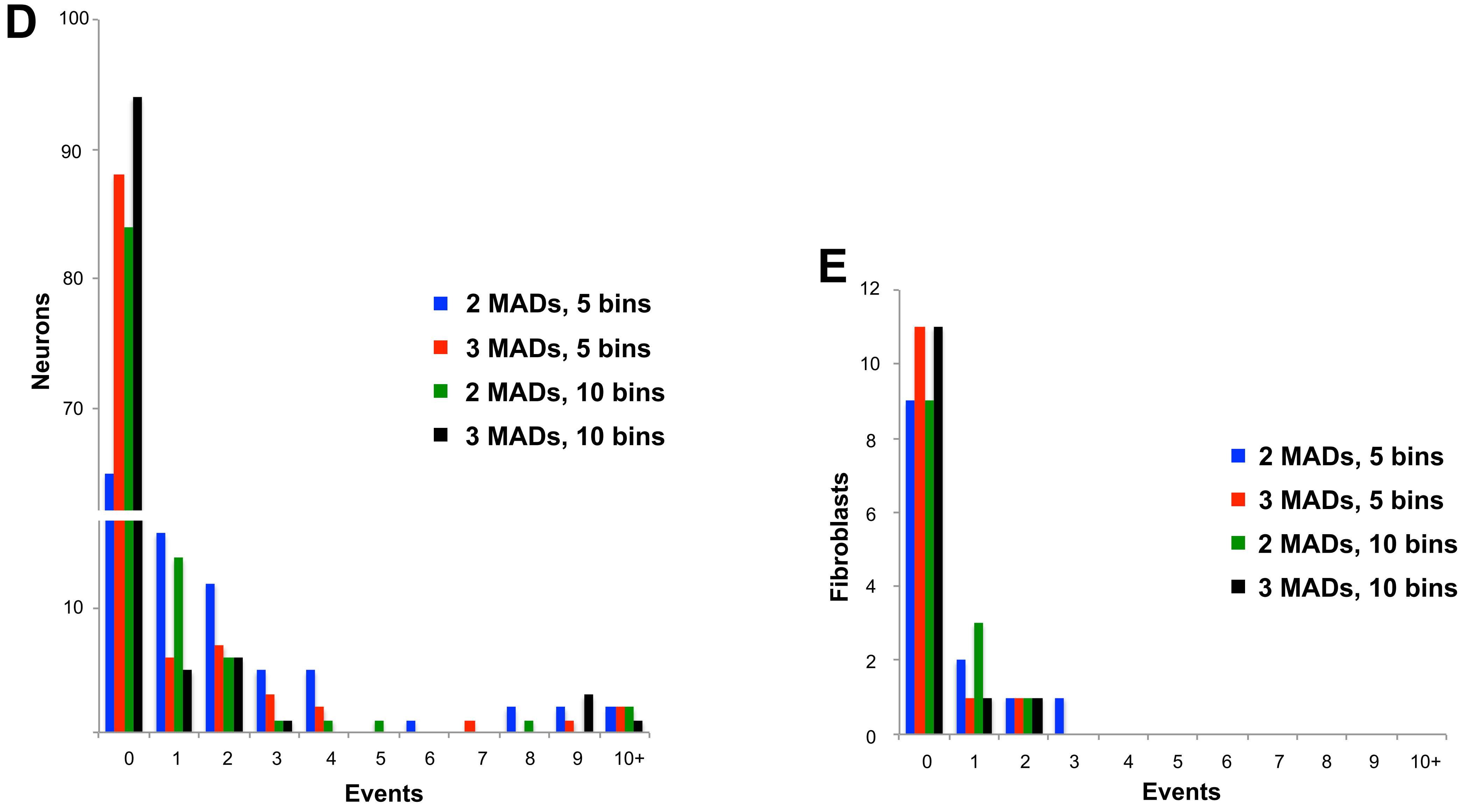


Figure S11. Effect of increased CNV calling stringency. (A) Table showing the effect of increasingly stringent CNV detection thresholds on the level and types of CNVs found in frontal cortex neurons. At far left we show the increasingly strict CNV filtering parameters where "MADs" refer to minimum amplitude of the CNV as measured by the number of median absolute deviations from the dataset median and "bin" refers to the minimum number of consecutive genomic windows identified by the segmentation algorithm. From left we show the total number of CNV calls detected, the number of deletions and duplications, the number and percentage of cells that were found to have at least 1 CNV, the predicted false negative rate (FNR) as calculated in the same manner as for Fig. S7 (see methods), and the fraction of male neurons in which monosomy X was detected at the given thresholds. Note that the false negative rate is calculated using simulated CNVs that are the same size as the minimum number of bins that could be detected according to the bin thresholds at far left (either 5 or 10), and therefore FNR actually decreases with the 10-bin threshold because larger CNVs are easier to detect. **(B)** The effect of increased stringency on the 13 control fibroblast cells (see Fig. S5A and D). In addition to the columns shown for part A, here we show the percentage of cells in which trisomy 21 was detected at the indicated thresholds. **(C)** The effect of increased stringency on the 3 single fibroblasts subjected to the replicate single cell sequencing experiment (see Fig. S5A-C). In addition to the columns described above, this table includes the number of concordant and discordant CNVs detected at each indicated threshold. Concordant CNVs are defined as those detected in both replicate cells; discordant CNVs are those detected in merely one replicate cell, according to the filtering thresholds shown at left. In one case two CNV calls in one replicate dataset were concordant with a single call in the pair, hence the odd number of concordant calls. **(D)** Bar chart showing the number of individual neurons (y-axis) that exhibited a given number of CNVs (x-axis) at the four CNV detection thresholds indicated in the legend. Note that the Y-axis is “broken”, with the section between ~18 and ~60 not shown. **(E)** Bar chart showing the number of fibroblasts (y-axis) that exhibited a given number of CNVs (x-axis) at the four CNV detection thresholds indicated in the legend. This plots is based on the 13 fibroblasts shown in part B and in Fig. S5A and D. See Table S3 for all fibroblast CNV calls.

Supplementary Tables: see separate files.

Table S1. The CNVs identified by single cell microarray experiments on hiPSC-derived neurons, hiPSC-derived NPCs, and donor fibroblasts. The first sheet is a key describing the columns.

Table S2. The CNVs identified by single cell sequencing experiments on fibroblasts originating from a male with trisomy 21. The first sheet is a key describing the columns. The second sheet contains CNV calls from single fibroblasts. The third sheet contains CNV calls in fibroblasts subjected to replicate single cell sequencing (see Fig. S5A). Note that there are a few cases where the segmentation algorithm called a CNV in only one of the two replicate datasets. In these cases we report the median and MAD values of the same genomic interval from the replicate pair in two additional columns (as described in the key). These numbers show that in cases of discordant CNVs, the apparent cause is false negative calls in the replicate pair that did not have a CNV call.

Table S3. The CNVs identified by single cell sequencing experiments on post-mortem frontal cortex neurons. The first sheet is a key describing the columns. Note that this table describes all raw CNV calls obtained by segmentation, and that the 137 subchromosomal CNVs described in the text does not include the multiple CNV calls that comprise 3 "whole chromosome" gains or losses, defined as cases where >50% of a single chromosome is affected by either gains or losses (but not both).